2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT) | 979-8-3503-8354-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/IC2PCT60090.2024.10486482

Flight Delay Prediction Using Machine Learning Techniques

Yash Tijil
Department of Computer Science & Engineering
Galgotias University
Greater Noida, India
yashtijil@gmail.com

Nripendra Dwivedi
Professor, Department of Computer Science & Engineering
Galgotias University
Greater Noida, India
nripendra.dwivedi@galgotiasuniversity.edu.in

Satyam Kumar Srivastava
Department of Computer Science & Engineering
Galgotias University
Greater Noida, India
satyamsrivastava7079530@gmail.com

Anmol Ranjan
Department of Computer Science & Engineering
Galgotias University
Greater Noida, India
anmolranjan1@gmail.com

***Abstract*—Flight delays critically impact passengers, airlines, and the economies of affected regions. We aimed to predict flight delays by developing a structured prediction system that utilizes flight data to forecast departure delays accurately. This project involved a comprehensive analysis of various machine learning methods, utilizing a dataset containing information related to flights. The primary focus was on extracting valuable insights from this extensive dataset to accurately predict flight delays. By conducting thorough assessments and comparative analyses, we appraised and contrasted these techniques regarding their efficacy in predicting flight delays to obtain valuable insights into the effectiveness of these methods. The methods suggested in this project are anticipated to provide airline companies with the ability to make accurate predictions of delays, improve flight planning, and reduce the impact of delays.**

Keywords—Machine learning, Flight Delay Prediction, Random Forest, Logistic Regression, Support Vector Machine (SVM)

I. Introduction

The economic impact of airline delays extends across the aviation sector, affecting airports, airlines, and passengers on a global scale [1]. The resulting operational disruptions lead to increased operational costs for airlines, reduced airport efficiency, and, consequently, diminished passenger satisfaction. The complexity of managing and mitigating the effects of delays necessitates a comprehensive understanding of the contributing factors. The Bureau of Transportation Statistics (BTS) estimates that delays account for 20% of all commercial flights [2].

This study incorporates multiple machine learning algorithms for a comparative analysis aimed at assessing the accuracy of each algorithm. The paper follows this structure: Section II entails a literature review derived from various sources, while Section III describes the techniques applied for data pre-processing and cleaning. Section IV provides an in-depth discussion of the methodologies employed and the comparative study conducted.

II. Literature Review

Many machine learning-based techniques in data mining like clustering, classification rules, and regression have been proposed to create and extract model predictions from historical data.

The configuration of the forecast model may indicate the lack of influence on the initiation of airport ground delay measures. For the purpose of weather-dependent analysis of dates and evaluation of performance, unsupervised data modelling approaches such as clustering are applied.

Akpinar and Karabacak used data mining categorization criteria, taking into account critical aspects such as airports, airlines, cargo, passengers, efficiency, and safety [3]. This study provides a comprehensive overview of data mining applications in civil aviation. Data mining may help with fuel price optimization, cargo optimization, passenger tracking, airport conditions, weather forecast, revenue per flight, cost per seat, catering and handling cost per seat etc.

Zhang and Nazeri used data mining to study the influence of weather on the performance of the National Airspace System (NAS) [4]. They use C5.0 decision tree learning technology and K-means clustering algorithm. The study found that weather patterns/conditions have an impact on NAS performance. They concluded that the revealed criteria pertain to flights that were blocked that can be used to predict performance on specific days based on weather.

Ha et al. used the CRISP-DM (CRoss Industry Standard Process for Data Mining) method to develop an experimental model and applied it to the exploration of big data [5]. They classified airports based on arrival delays and deduced that they could make recommendations on the best airports for arrival delays.

Sridhar and Mukherjee utilized two models: decision tree and logistic regression to predict when a Ground Delay Program will

take place relying on traffic demand and climatic circumstances [6]. These algorithms are used to estimate the GDP of two major US airports. Evaluation of the models is done using data from weather variables such as wind, convection, precipitation, cloud height and visibility, as well as arrival requirements based on flight time. The logistic regression method calculates the probability of GDP occurring in that period, while the decision tree classifies hours as GDP or non-GDP.

Natarajan et al. utilized logistic regression and decision trees (random forest) using the same method to estimate delays [7]. They also evaluated the projected arrival time and delays for both models and determined that the decision tree method was more successful.

Tu et al. used probabilistic models to demonstrate that when delays are smaller than 2 hours, the chance of delay may be predicted [8]. Mueller et al. evaluated Normal and Poisson distributions and concluded that although arrival delays can be fitted to a normal distribution, the Poisson distribution is appropriate for measuring departure delays [9].

A review of several methods for forecasting and assessing flight delays was conducted by Sternberg et al. Mueller et al. discovered, in contrast, that delayed arrivals follow a Poisson distribution, but late departures follow a normal distribution [10].

Lu conducted a study on flight delays using Bayesian network and decision tree models and came to the conclusion that it is difficult to anticipate aircraft delays with accuracy [11].

The outcomes of neural networks, decision trees, and Naïve Bayes were also assessed in a study by Lu et al. where decision trees performed the best, with a prediction confidence of 70% [12].

As per Chen et al., fuzzy SVM with a weighted margin is more accurate for predicting flight delays than a standalone SVM [13].

Delay predictions have also been made using ST-Random Forest and CNN-LSTM deep learning frameworks with 92.39% accuracy [14, 15].

III. Methodology

Only when a flight arrives fifteen minutes or more later than expected is it classified as delayed by the BTS [2]. Any flight that is running more than fifteen minutes behind schedule is labelled as ‘Delayed’ in this article. In order to increase accuracy and yield better results those entries that have cancelled or diverted flights have been excluded. Another column called ‘Delay’ has been obtained using the data from the ‘Departure Delay’ column. There are two values in the ‘Delay’ column: 0 and 1. The flight status is represented by these values. The numbers 0 and 1 correspond to the flights that took off on time and those that left beyond 15 minutes after the planned departure time. Next, we used the ‘Delay’ column to determine the f1-score, accuracy, recall, and support.

In this section, we provide an overview of the suggested framework. The most crucial stage in creating a model is data collection, which comes first. When obtaining records, it is essential to consider important factors, including the validity, accuracy and legality of the dataset. Data collection was followed by the data pre-processing stage, during which the acquired data was cleaned up and formatted. Data cleaning was also performed during this phase, and it included eliminating null values from tuples, eliminating unnecessary information, etc. We added ‘Delay’ as a new column for efficiently handling the data. The entire data was divided into two groups in this column: delayed and on time. This step produced data that was processed to ensure that the algorithms could use it. In this approach, support vector machine (SVM), logistic regression and random forest algorithms were used. After analyzing the dataset, pertinent characteristics were extracted to prepare for the testing and training stages. Each of the three algorithms went through a rigorous testing and training process. This was followed by a comparison of the three algorithms' respective f-score, recall, precision, and accuracy scores.

IV. Methods

A. Random Forest

Random forest is a supervised machine learning model that can be utilized in classification as well as regression tasks. It integrates several classifiers to address intricate issues and improve the model's performance [16]. The construction of random forest algorithms involves the aggregation of predictions from various decision trees, each individually trained.

B. Support Vector Machine

Algorithms for supervised learning in regression and classification applications are called SVMs. SVM aims to maximize the geometric margin while minimizing the empirical classification error [17]. To achieve classification and regression, this entails building hyperplanes in infinite space or high dimensional space. For a set of valid points, every labelled member of a set of practice data points builds a model that can be used to classify new examples into one of these categories. Because of this feature, SVM is a non-linear binary classifier.

The optimization problem is solved to derive the parameters of the maximum-margin hyperplane.

C. Logistic Regression

It is a statistical technique employed to estimate the probability of a binary outcome. Despite its name, classification tasks are the ones for which logistic regression is used, not regression tasks. It belongs to the general linear model (GLM) category and represents the association between one or more independent variables and a binary dependent variable by estimating probabilities.

V. Data Procurement and Preprocessing

A. Dataset

The Bureau of Transportation Statistics (BTS) was used as the data source for acquiring the following comprehensive flight information data for the year 2015:

‘Month,’ ‘Day,’ ‘Day of Week,’ ‘Year,’ ‘Flight Number,’ ‘Airline,’ ‘Tail Number,’ ‘Origin Airport,’ ‘Destination Airport,’ ‘Scheduled Departure,’ ‘Departure Time,’ ‘Departure Delay,’

'Taxi Out,' 'Wheels Off,' 'Scheduled Time,' 'Elapsed Time,' 'Air Time,' 'Distance,' 'Wheels On,' 'Taxi In,' 'Scheduled Arrival,' 'Arrival Time,' 'Arrival Delay,' 'Diverted,' 'Cancelled,' 'Cancellation Reason,' 'Air System Delay,' 'Airline Delay,' 'Aircraft Delay,' 'Aircraft Delay,' and 'Weather Delay.'

B. Data Preprocessing

Prior to model training, it is essential to preprocess the data to prevent potential errors later on. In this study, various Python programming techniques and libraries were used to preprocess the data:

1. Columns with reasons for delay had some null values. They were replaced with 0.

2. Handling missing values: The 'Arrival Delay' and 'Departure Delay' columns had some missing values. These rows were dropped.

3. Eliminating superfluous traits: While the majority of the traits are pertinent, some were discarded since they weren't necessary.

For example, in our dataset, the 'Cancelled' column was removed because cancelled flights are not considered delayed in this paper; the associated column 'Cancellation Reason' was removed as well. The column 'Diverted Flights' was dropped for the same reason. Categorical columns such as 'Airline,' 'Tail Number,' 'Origin Airport,' and 'Destination Airport' were removed as the utilized models do not accept categorical variables.

VI. Results

A. Confusion Matrix

After training, the model was tested and obtained an approximate 100% Accuracy Score. This suggests that every piece of examined data was correctly classified by the machine learning model. Within the confusion matrix, the diagonal components match the total number of correctly classified tuples [1]. Here is a summary of the Confusion Matrix:

```
[[ 41988,     0],
 [     0, 159113]]
```

Fig. 1. Confusion Matrix generated for Support Vector Machine model

```
Accuracy: 1.0
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    159113
           1       1.00      1.00      1.00     41988

    accuracy                           1.00    201101
   macro avg       1.00      1.00      1.00    201101
weighted avg       1.00      1.00      1.00    201101
```

Fig. 2. Accuracy, recall, f1-score, precision, and support for the support vector machine model

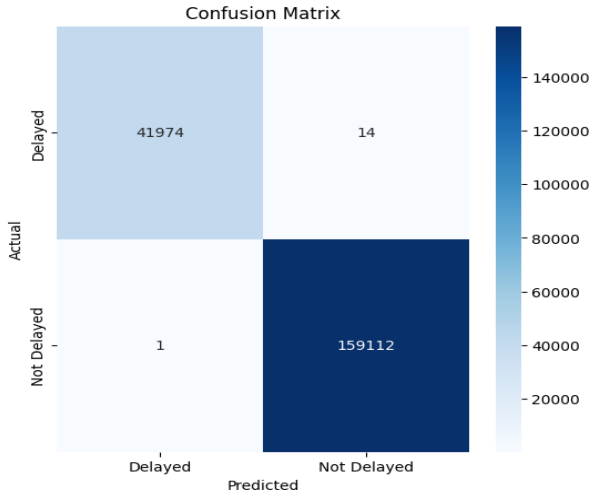

Fig. 3. Confusion matrix of Logistic Regression model

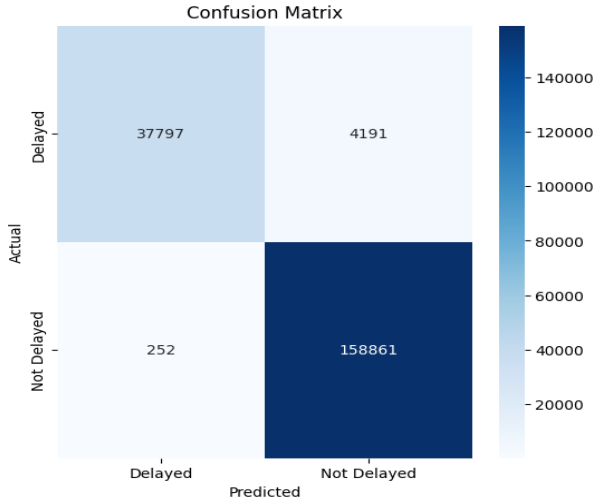

Fig. 4. Confusion matrix of Random Forest Model

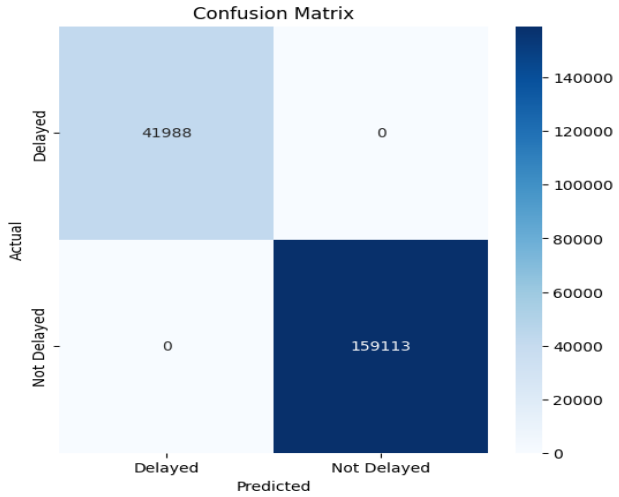

Fig. 5. Confusion matrix of Support Vector Machine model

B. Comparative Analysis of Employed Algorithms

The models were applied to the dataset with the main purpose of identifying and classifying flights with delays of more than 15 minutes. The 'Delay' attribute within the dataset served as the target variable for assessing and classifying flight delays.

Classification of Delays: The model training incorporated parameters such as 'Flight Number,' 'Air Time,' 'Taxi Out,' and 'Weather Delay,' utilizing the 'Delayed' attribute within the dataset. Subsequently, various algorithms were applied to the dataset, yielding the following sample counts:

- There are 804,412 training samples.
- 201,101 testing samples in total.

The model's functionality was assessed using the following criteria:

- Validation accuracy characterizes the model's ability to accurately predict samples within a set of given values.
- Recall is determined by the ratio of the correct return of critical events to all related events.
- True positive / (true positive + false negative) equals recall.
- Precision is determined by the ratio of accurately identified positive instances to the total positive observations.
- True positive / (true positive + false positive) equals precision.
- F1-Score or weighted average of recall and precision, is a statistically defined accuracy metric.

TABLE I. CONDENSED RESULTS OF F1-SCORE, ACCURACY, RECALL, AND PRECISION FROM THE ALGORITHMS

Algorithm	Precision		Recall		F1-Score		Accuracy
	0	*1*	*0*	*1*	*0*	*1*	
Random Forest	0.97	0.99	1.00	0.90	0.99	0.94	0.97
Logistic Regression	1.0	1.0	1.0	1.0	1.0	1.0	0.99
SVM	1.00	1.00	1.00	1.00	1.00	1.00	1.0

- Therefore, achieving 100% accuracy with the SVM learning model when provided with any set of attributes (feature values) distinctly demonstrates its efficiency for predicting aircraft delays.

VII. CONCLUSION

Our research showed that aircraft delay predictions may be made with good accuracy using machine learning methods. In addition to evaluating delays for different human requirements, the previously mentioned categorization and analysis aims to closely examine elements that impact delays, including 'weather delay,' 'airline delay,' and 'security delay.' The SVM model performed the best, correctly categorizing delays into two groups using the previously mentioned parameters including 'Departure Time,' 'Air Time,' and 'Month'—with 100% accuracy. Thus, it can be effectively used to predict flight delays, which will be beneficial for airports and airlines, as well as passengers.

Therefore, the study of flight delays presented in this paper is grounded entirely in scientific parameters, underscoring its crucial significance in the aviation industry.

A. Future Scope

Although this study examined several elements potentially impacting flight delays, future research can focus on analyzing the effect of other factors such as temporal patterns and seasonality. Furthermore, deep learning models, such as recurrent neural networks, can be utilized to capture sequential patterns in the data.

REFERENCES

[1] Borse, Y., Jain, D., Sharma, S., Vora, V., and Zaveri, A. (2020) Flight Delay Prediction System. International Journal Of Engineering Research & Technology (IJERT) Volume 09, Issue 03 (March 2020).

[2] Bureau of Transportation Statistics. Available online: https://www.bts.gov/ (accessed on 26 March 2020).

[3] Akpinar, M.T. and Karabacak, M.E. (2017). Data mining applications in civil aviation sector: State-of-art review. In *CEUR Workshop Proc* (Vol. 1852, pp. 18-25).

[4] Nazeri, Z. and Zhang, J. (2017). Mining Aviation Data to Understand Impacts of Severe Weather. In *Proceedings of the International Conference on Information Technology: Coding and Computing* (ITCC.02) pp. 518-523.

[5] Ha. ,. J. N. a. H. P. S. Man. (2015) "Analysis of Air-Moving on Schedule Big Data based on CrispDm Methodology," ARPN Journal of Engineering and Applied Sciences, pp. 2088-2091.

[6] Mukherjee, A., Grabbe, S. R., and Sridhar, B. (2014). Predicting Ground Delay Program at an airport based on meteorological conditions. In 14th AIAA Aviation Technology, Integration, and Operations Conference (pp. 2713-2718).

[7] Natarajan, V., Meenakshisundaram, S., Balasubramanian, G. and Sinha, S. (2018) "A Novel Approach: Airline Delay Prediction Using Machine Learning," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA pp. 1081-1086, doi: 10.1109/CSCI46756.2018.00210

[8] Tu, Y., Ball, M.O, and Jank, W.S. (2008) Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern", Journal of the American Statistical Association, 103, pp. 112-125.

[9] Mueller, E.R. and Chatterji, G.B. (2002) Analysis of aircraft arrival and departure delay characteristics", In AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum, Los Angeles, California, USA.

[10] Sternberg, A., Soares, J., Carvalho, D., and Ogasawara, E. (2017) A review on flight delay prediction. arXiv preprint arXiv:1703.06118.

[11] Lu, Z. (2010) Alarming Large Scale of Flight Delays: An Application of Machine Learning, Machine Learning. In Tech publishing, pp. 239-250.

[12] Lu, Z., Wang, J., and Zheng, G. (2008) A new method to alarm large scale of flights delay based on machine learning, in knowledge acquisition and modelling", KAM '08. International Symposium on, pp. 589-592.

[13] Chen, H., Wang, J., and Yan, X. (2008) A fuzzy support vector machine with weighted margin for flight delay early warning", In *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery* (Vol. 3, pp. 331-335). IEEE.

[14] Li, Q. and Jing, R. (2022) Flight delay prediction from spatial and temporal perspective, Expert Systems with Applications, Volume 205, 117662.

[15] Li, Q., Guan, X., and Liu, J. (2023) A CNN-LSTM framework for flight delay prediction, Expert Systems with Applications. Volume 227, 120287.

[16] Leo Breiman - Random Forests (Dept. of statistics, University of California, Berkeley)

[17] Durgesh K.Srivastava, Lekha Bhambu – Data Classification using Support Vector Machine